



United States Department of Agriculture

# The i5k Workspace@NAL

*A Tripal based Arthropod genome portal*

Christopher Childers  
USDA/ARS/NAL  
[i5k.nal.usda.gov](http://i5k.nal.usda.gov)

# Background

- The i5k initiative tasked itself with coordinating the sequencing and assembly of 5000 insect or related arthropod genomes
- The i5k Workspace@NAL is available to help any i5k project with genome hosting needs
  - Support community needs
  - Connect researchers to the data
  - Create standardized tools for accessing the data in useful ways
  - Provide resources to facilitate annotation projects

# The i5k Workspace@NAL

- Diverse representation across Arthropoda
  - 55 species and counting
  - 2 had OGS when they came to us
  - 8 have completed an annotation phase and have OGS
    - 3 from external collaborator
    - 5 from our workflow
    - 2 in process
  - 47 species are accepting applications for annotators
- Facilitate collaboration between groups
  - Annotators work on related taxa
  - Domain specialists that focus on gene families across multiple species



United States Department of Agriculture

'Frozen' genome  
assembly

Automated  
annotations

Ancillary datafiles (e.g.  
RNA-Seq alignments)

Submission

 **Workspace@NAL**  
<https://i5k.nal.usda.gov/>

## Resources

## Tools

## Services

### Organism Information Page

### Custom BLAST interface

Manual annotation  
quality control

Official gene set generation

### Bulk data downloads

### JBrowse genome browser

## Challenges

Non-standard data  
formatting

### Tutorials

### Apollo manual curation tool

Failure to submit all metadata  
(ex: sample origin; analysis  
methods)

HMMer

Clustal



United States Department of Agriculture



United States Department of Agriculture  
National Agricultural Library

[Organisms](#) ▾ [Data](#) ▾ [Tools](#) ▾ [Tutorials and Resources](#) ▾ [Contact](#) [About us](#)

[Login](#)

## i5k Workspace@NAL

*A place for arthropod genome communities to curate, visualize and share data*

Search



© Scott Bauer

[Source link](#)

### Apollo/JBrowse

[View guidelines](#)

Manually curate a genome with Apollo, or browse a genome and its features with JBrowse.

[REGISTER](#)

### BLAST

[View Tutorial](#)

Search all available genomes and gene sets with BLAST

[RUN BLAST](#)

## Join an i5k Workspace Project

Follow the instructions to join one or more manual annotation projects



[Read our annotation guidelines](#)



[Register for access to the annotation system](#)



[Begin annotating!](#)

## Start an i5k Workspace Project or Submit Data

We are happy to host any arthropod genome project. [Learn more about sharing your genome project or dataset.](#)

[Submit Data](#)

# Organism Pages



United States Department of Agriculture  
National Agricultural Library

i5k Workspace@NAL

Organisms ▾ Data ▾ Tools ▾ Tutorials and Resources ▾ Contact About us


[Login](#)

[Organisms](#) / [Catajapyx aquilonaris](#)

## Catajapyx aquilonaris

[Overview](#)

[Annotation Methods](#)

[Assembly Methods](#)

[Catajapyx aquilonaris @ Baylor College of Medicine](#)

[NCBI BioProject](#)

### Overview



The japygid *Catajapyx aquilonaris* is a blind predator of the soil. Like Protura (*Acerentomon maius*) and Collembola (*Sminthurus viridis*), Diplura lack wings, mirroring the wingless insect ancestor. Like in all primarily wingless hexapods, sperms are not transferred directly during copulation. Males rather deposit a spermatophore on the ground and females subsequently take the spermatophore up.

Diplura are critical for understanding the evolutionary origin of Hexapoda (e.g., terrestrialization), the evolutionary origin of wings (ancestral condition in Diplura), and the evolution of direct sperm transfer (ancestral condition in Diplura).

Data were generated by the [Baylor College of Medicine's i5k pilot project](#).

[View the Baylor College of Medicine's data sharing policy.](#)

Image Credit: Copyright Nikola Szucsich

### Catajapyx aquilonaris data files

Name	Last modified
<a href="#">← Parent Directory</a>	
<a href="#">Current Genome Assembly</a>	2015-03-19 13:42

Assembly Information

Statistics



# Organism Pages

Image Credit: Copyright Nikola Szucsich

## Assembly Information

Analysis Name	Whole genome assembly of Catajapyx aquilonaris
Software	Baylor College of Medicine genome assembly pipeline (NA)
Source	<a href="#">forcepstail.consistent.scaffolds</a>
Date performed	2014-10-07
Materials & Methods	<p>Sequence generation for assembly. For this project we are generating fairly high coverage in a number of different insert sized libraries. The assembly strategy is based around a seed allpaths assembly (the Broad Allpaths assembler) followed by seed assembly improvement using homegrown tools, <a href="#">Atlas-link</a> and <a href="#">Atlas-GapFill</a>, which can significantly improve the results. Thus we generate sequence data to enable the Allpaths assembly. As of Nov 2011 this is: - 40X genome coverage in 180bp insert library (100bp reads forward and reverse); and 40X 3kb insert data. To enable better scaffolding and local gap filling we additionally generate 500bp, 1kb, 2kb, and 8kb insert sizes at &gt; 20X coverage.</p> <p><a href="#">Source: Baylor College of Medicine i5K Project Summary</a></p>

## Statistics

Assembly Metrics	
Contig N50	11472
Scaffold N50	30909
GC Content	44.41
Manual Annotations	



# Gene Pages



## Dicer-2, OFAS025276 (gene) *Oncopeltus fasciatus*

### Overview

### Sequences

### Transcripts

#### Overview

**Organism** *Oncopeltus fasciatus*  
**Gene ID** OFAS025276  
**Gene Name** Dicer-2  
**Synonyms** NA  
**Location** Scaffold23:319420..445740+  
**Transcripts** This gene contains [1 mRNA](#)  
**Analysis** *Oncopeltus fasciatus* Official Gene Set v1.1  
Source: *Whole genome assembly of Oncopeltus fasciatus*  
**Annotator Comments** None

#### Available Tracks

✕ filter by text

##### 0. Reference Assembly 2

- ☐ GC Content
- ☐ Gaps in assembly

##### 1. Official Gene Set v1.2 4

###### 1. Gene Sets 4

- ▶ Noncoding 1
- ▶ Other features 1
- ☐ OGSv1.2 sequence modifications



File View Help

Full-screen view

Login



Scaffold23

Scaffold23:319454..445774 (126.32 K)

Go



350,000

400,000

OGSv1.2 protein-coding genes

DNA

ORF

Dicer-2 - part 1 of 2

Note to curator: complete, concatenated CDS:...

# Annotator Registration

- Registration through a custom in-house module
- All registrants are reviewed before approval
- Permissions are at the organism level

## Web Apollo registration form

Complete the form below and click 'Submit' to register for a Web Apollo account. Only registered users can view, create or change annotations.

Full Name \*

Email Address \*

Organism \*

Select one or multiple organisms

Agrilus planipennis

Athalia rosae

Blattella germanica

Institution \*

Genes or gene families that you intend to annotate \*

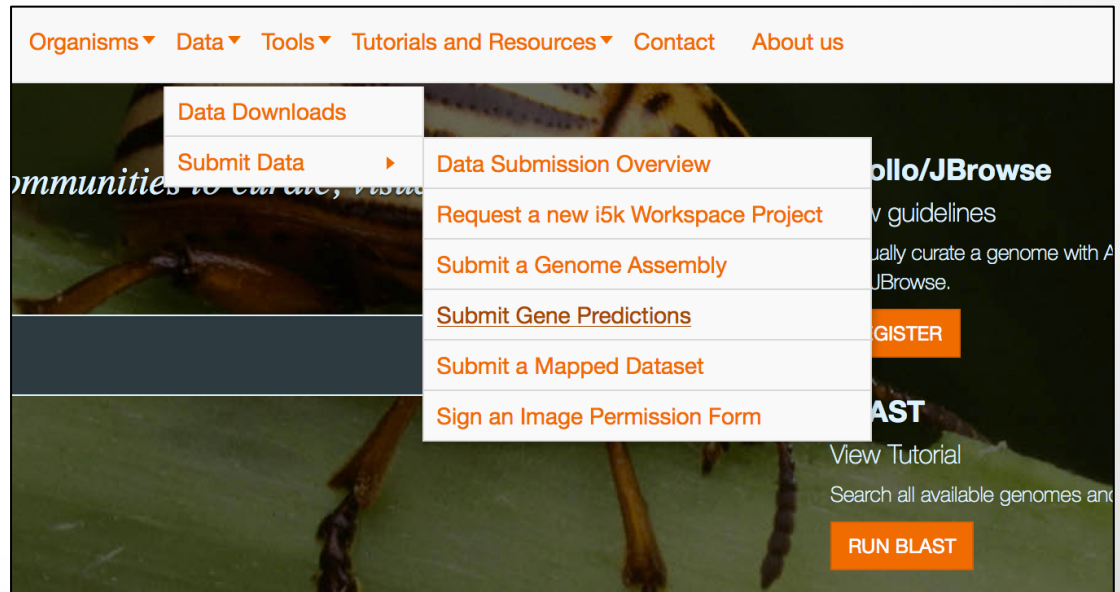
Math question \*

2 + 7 =

Solve this simple math problem and enter the result to help us reduce spam. E.g. for 1+3, enter 4.

# Data Submission

- Dynamic web forms
  - Standardized and validated metadata inputs
- Replaces the spreadsheet submission system
- Under the new system:
  - 4 new species
  - 4 assemblies
  - 4 gene prediction sets
  - 2 mapped datasets





United States Department of Agriculture

'Frozen' genome  
assembly

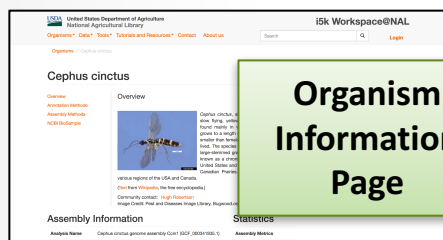
Automated  
annotations

Ancillary datafiles (e.g.  
RNA-Seq alignments)

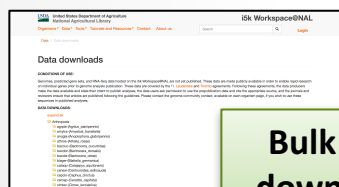
Submission

 **Workspace@NAL**  
<https://i5k.nal.usda.gov/>

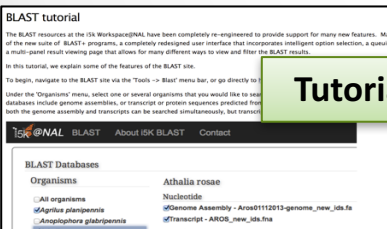
## Resources



Organism  
Information  
Page

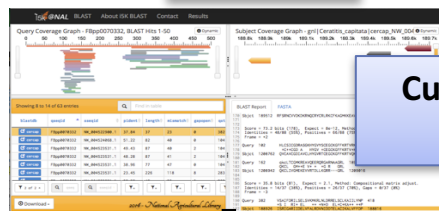


Bulk data  
downloads

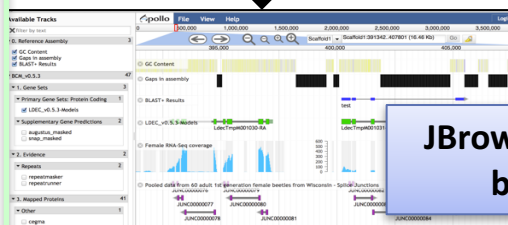


Tutorials

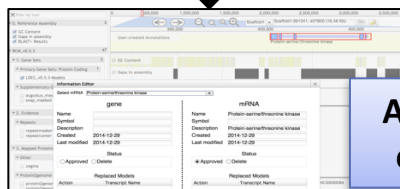
## Tools



Custom BLAST  
interface



JBrowse genome  
browser



Apollo manual  
curation tool

HMMer

Clustal

## Services

Manual annotation  
quality control

Official gene set generation

## Challenges

Non-standard data  
formatting

Failure to submit all metadata  
(ex: sample origin; analysis  
methods)

# Search/Alignment tools

- Web interface is built in Python/Django
  - Interface dynamically updates options based on inputs
  - Results persist for one week
- Clustal
  - Multiple sequence alignment package
  - We provide both Clustalw and Clustal-Omega
  - Results may be passed directly to HMMer
- HMMer
  - Sequence search tool more sensitive than BLAST
  - Uses probabilistic models to find distant homologs
  - Can accept single sequences (FASTA) or Multiple alignments (MSA)
  - Available for use with our protein sets
- BLAST
  - Sequence alignment tool
  - Search across genome, CDS or peptide databases
  - Submit multiple sequences at once
  - Search multiple databases with one submission



United States Department of Agriculture

# BLAST

### BLAST Databases

#### Organisms

- ☐ All organisms
- ☐ *Agrilus planipennis*
- ☐ *Anoplophora glabripennis*
- ☐ *Athalia rosae*
- ☒ *Bactrocera cucurbitae*
- ☐ *Bactrocera dorsalis*
- ☒ *Blattella germanica*
- ☐ *Cataglyphis aquilonaris*
- ☐ *Centruroides exilicauda*
- ☐ *Ceratitis capitata*
- ☒ *Cimex lectularius*
- ☐ *Copidosoma floridanum*
- ☐ *Dianhorina citri*

#### Query Sequence

Your sequence is detected as peptide:  
WVESAGNVISLQGFSYNAVHFALCHIYSGA  
SNIPETINIVELATLADMLCLEGLKEVIMYTL  
KVYKCHFFHKPCNSCISGVLECLPLAAA  
YGLDEIYKSLRWITKYFVRVWPTKGFANL  
PKELQDKCYQHIVHLSAENVLETIMGCEK  
LEATVNVKWAQTVINMNLKHEASVKYL  
TQHFADVLSSEA  
Or load it from disk  
 No file selected.

#### Program

☐ blastn ☒ tblastn ☐ tblastx ☐ blastp ☐ blastx

#### Cimex lectularius

##### Nucleotide

- ☒ Genome Assembly - Clec\_Bbug02212013.genome\_new\_ids.fa
- ☒ Transcript - CLEC\_new\_ids.fna

##### Peptide

- ☐ Protein - CLEC\_new\_ids.faa

#### Query Coverage Graph - CLEC000004-PA, BLAST Hits 1-50

#### Subject Coverage Graph - gnl | Cimex\_lectularius | cimlec\_Scaffold1, BLAST Hits 1-50

#### Showing 1 to 11 of 51 entries

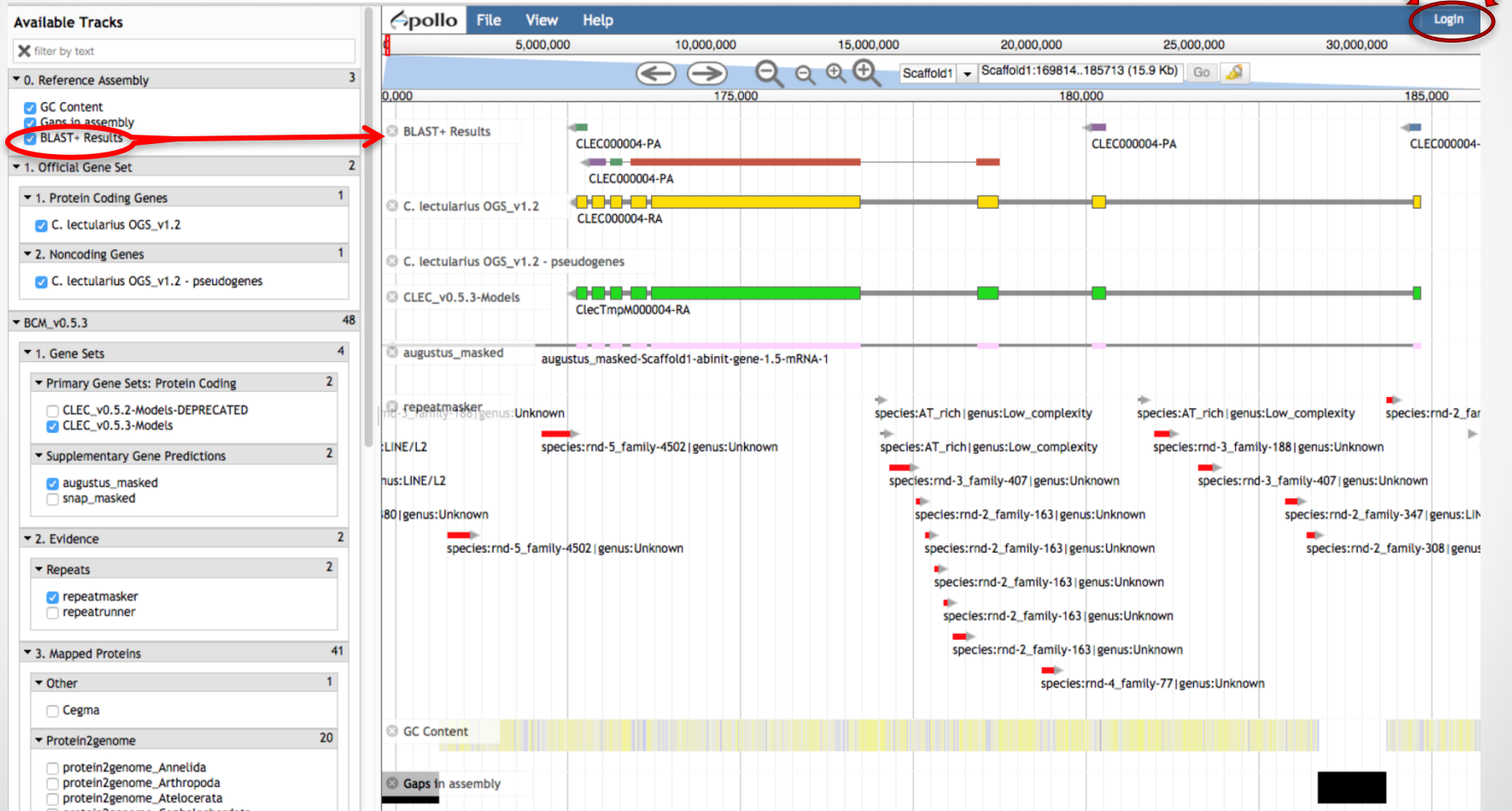
blastdb	qseqid	sseqid	pident	length	mismatch	gapopen	qseq
cimlec	CLEC000004-PA	CLEC000004-RA	100	1480	0	0	1
cimlec	CLEC000004-PA	Scaffold1	98.29	1111	0	1	2
cimlec	CLEC000004-PA	Scaffold1	85.37	82	11	1	1
cimlec	CLEC000004-PA	Scaffold1	100	59	0	0	1
cimlec	CLEC000004-PA	Scaffold1	92.86	112	8	0	1
cimlec	CLEC000004-PA	Scaffold1	100	69	0	0	4
cimlec	CLEC000004-PA	Scaffold1	100	55	0	0	1
cimlec	CLEC000004-PA	Scaffold1	80	55	9	1	1
baccuc	CLEC000004-PA	NW_011863740.1	58.6	186	72	2	1
baccuc	CLEC000004-PA	NW_011863740.1	30.69	202	97	10	6

#### BLAST Report

##### FASTA

263 Sbjct 172874 KCYKQIIVHL-VHNTFSLIKSC 172812  
264  
265  
266 Score = 130 bits (326), Expect(2) = 2e-68, Method: Compositional matrix adjust.  
267 Identities = 59/59 (100%), Positives = 59/59 (100%), Gaps = 0/59 (0%)  
268 Frame = -1  
269  
270 Query 1301 FSYNAVHFALCHIYSGASNPETINIVELATLADMLCLEGLKEVIMYTLKVYKCHFFHK 1359  
271 FSYNAVHFALCHIYSGASNPETINIVELATLADMLCLEGLKEVIMYTLKVYKCHFFHK  
272 Sbjct 173287 FSYNAVHFALCHIYSGASNPETINIVELATLADMLCLEGLKEVIMYTLKVYKCHFFHK 173111  
273  
274  
275 Score = 213 bits (542), Expect = 3e-53, Method: Compositional matrix adjust.  
276 Identities = 104/112 (93%), Positives = 106/112 (95%), Gaps = 0/112 (0%)  
277 Frame = -3  
278  
279 Query 102 LIESFIREVYTRNVKHKEDAVVAQIKIVSSSKEVESPDSDVFVTPKASCPCEIPNTRQ 161  
280 +I REVYTRNVKHKEDAVVAQIKIVSSSKEVESPDSDVFVTPKASCPCEIPNTRQ  
281 Sbjct 178575 MIFPVCREVYTRNVKHKEDAVVAQIKIVSSSKEVESPDSDVFVTPKASCPCEIPNTRQ 178578  
282  
283 Query 162 TALTPASKRNEINIDELIKRHHYNNLVHVEEDLFKELAEQDALSNSFRSIL 213  
284 TALTPASKRNEINIDELIKRHHYNNLVHVEEDLFKELAEQDALSNSFRSIL  
285 Sbjct 178577 TALTPASKRNEINIDELIKRHHYNNLVHVEEDLFKELAEQDALSNSFRYVF 178422  
286  
287  
288 Score = 150 bits (378), Expect = 5e-34, Method: Composition-based stats.

# Feature Visualization





# Annotation via Apollo

**Available Tracks**

✕ filter by text

▼ 0. Reference Assembly 4

- ☐ Contamination
- ☒ GC Content
- ☒ Gaps in assembly
- ☒ BLAST+ Results

▼ 1. Official Gene Set 2

▼ 1. Protein Coding Genes 1

- ☒ C. lectularius OGS\_v1.2

▼ 2. Noncoding Genes 1

- ☒ C. lectularius OGS\_v1.2 - pseudogenes

▼ BCM\_v0.5.3 48

▼ 1. Gene Sets 4

▼ Primary Gene Sets: Protein Coding 2

- ☐ CLEC\_v0.5.2-Models-DEPRECATED
- ☒ CLEC\_v0.5.3-Models

▼ Supplementary Gene Predictions 2

- ☒ augustus\_masked
- ☐ snap\_masked

▼ 2. Evidence 2

▼ Repeats 2

- ☐ repeatmasker
- ☐ repeatrunner

▼ 3. Mapped Proteins 41

▼ Other 1

- ☐ Cegma

▼ Protein2genome 20

- ☐ protein2genome\_Annelida

apollo File View Tools Help

5,000,000 10,000,000 15,000,000 20,000,000 25,000,000 30,000,000

Scaffold1 Scaffold1:165601..186860 (21.26 Kb) Go

170,000 175,000 180,000 185,000

User-created Annotations

CLEC000004-RA

BLAST+ Results

CLEC000004-RA

CLEC000004-RA

C. lectularius OGS\_v1.2

CLEC000004-RA

C. lectularius OGS\_v1.2 - pseudogenes

CLEC\_v0.5.3-Models

ClecTmpM000004-RA

augustus\_masked

augustus\_masked-Scaffold1-abinit-gene-1.5-mRNA-1

GC Content

Gaps in assembly



# Clustal

## Clustal Omega

CLUSTALO CLUSTALW Tutorial

**Query Sequence**

Your sequence is detected as peptide:

>gi|913322868|ref|XP\_013192675.1| PREDICTED: N-alpha-acetyltransferase 15, NAtA auxiliary subunit-like, partial [Amyelosis transietella]

MPSPNSLPPEKALFRKLLCYEHQYKNGLKFAKQLSNPKFAEHGETLAMKGLTNLGRDEAYEVYKGLRND  
LKSPVCWHVYGLLRQSDKHQYEAICYNALKWEKNQILRLDLLQIMQRLDQYKDTROYQLMRPTQASWIG  
FAMSYHLGDYEMANSILDAFRNTQMKGPYDYEHSLLYQNMVLAESGGYERALQHLHKQSQILDLSSKETSG  
EYLLKLRKFEAEVYDLKRNPENVMYHKLIEAKQLVPEDEKVAFFDIYKKEYPRAIPRRLQTEARQDAFGRDVD  
DYLRHGLHGKIPPLVDLRSYVDQSADTIEKLQVLENLSKTGFSADASEWQPASALLWAYYYAQHFDYKKDT  
DRAHLYDAIEHTPTLLELPIYGRYKHAQDPSAYQWLEAQVMQADRYNYSKCAKRYMLAGHYKQAEQDMCAKF  
TREGVATENLNMQCMWQTALALAYQRTQQWQSEALKAHVEYDFHSEMEDQDFHSCYCMKMTLRSYVGLL  
RLEDVLRSHPYFRCAVIAQVRLRYACPLQAVQATTEPDGTGNFLFLHSEIMEDQDFHSCYCMKMTLRSYVGLL  
RLEDVLRSHPYFRCAVIAQVRLRYACPLQAVQATTEPDGTGNFLFLHSEIMEDQDFHSCYCMKMTLRSYVGLL  
->gml|Drosophila\_bipunctata\_protein\_v0.5.3| DBIP012051-PA

Or load it from disk

Browse... No file selected.

**Sequence Input**

Design input sequences

☒ yes

☐ no

Reset

Search

**Clustering and Iteration**

MBED-Like Clustering Guide-Tree

☒ yes

☐ no

MBED-Like Clustering Iteration

☒ yes

☐ no

**Iteration**

NUMBER OF COMBINED ITERATIONS

0

MAX GUIDE TREE ITERATIONS

Default

MAX HMM ITERATIONS

Default

**Output**

Format

Clustal

Out Order

Aligned

OR

## ClustalW

CLUSTALO CLUSTALW Tutorial

**Query Sequence**

Your sequence is detected as peptide:

>gml|Drosophila\_bipunctata\_protein\_v0.5.3| DBIP012051-PA

MPSPNSLPPEKALFRKLLCYEHQYKNGLKFAKQLSNPKFAEHGETLAMKGLTNLGRDEAYEVYKGLRND  
LKSPVCWHVYGLLRQSDKHQYEAICYNALKWEKNQILRLDLLQIMQRLDQYKDTROYQLMRPTQASWIG  
FAMSYHLGDYEMANSILDAFRNTQMKGPYDYEHSLLYQNMVLAESGGYERALQHLHKQSQILDLSSKETSG  
EYLLKLRKFEAEVYDLKRNPENVMYHKLIEAKQLVPEDEKVAFFDIYKKEYPRAIPRRLQTEARQDAFGRDVD  
DYLRHGLHGKIPPLVDLRSYVDQSADTIEKLQVLENLSKTGFSADASEWQPASALLWAYYYAQHFDYKKDT  
DRAHLYDAIEHTPTLLELPIYGRYKHAQDPSAYQWLEAQVMQADRYNYSKCAKRYMLAGHYKQAEQDMCAKF  
TREGVATENLNMQCMWQTALALAYQRTQQWQSEALKAHVEYDFHSEMEDQDFHSCYCMKMTLRSYVGLL  
RLEDVLRSHPYFRCAVIAQVRLRYACPLQAVQATTEPDGTGNFLFLHSEIMEDQDFHSCYCMKMTLRSYVGLL  
RLEDVLRSHPYFRCAVIAQVRLRYACPLQAVQATTEPDGTGNFLFLHSEIMEDQDFHSCYCMKMTLRSYVGLL  
->gml|Drosophila\_bipunctata\_protein\_v0.5.3| DBIP012051-PA

Or load it from disk

Browse... No file selected.

**Basic**

Sequence Type

☒ DNA

☐ Protein

Pairwise Alignment

☒ Full

☐ Fast

Reset

Search

**Full options**

Protein weight matrix

Default

Gap Open Penalty

Gap Extension Penalty

**Multiple Alignment**

Protein weight matrix

Default

Gap Opening Penalty

Gap Extension Penalty

Gap Separation Penalty

Output None

Maximum Number of Iterations

**Output**

Format

Clustal

Out Order

Aligned

### CLUSTAL Success

#### Download

Alignment

Submission Details

#### Report Details

CLUSTAL O(1.2.0) multiple sequence alignment

```
gml|Eurytemora_affinis_protein_v0.5.3|EAFP027431-PA
gml|Tigriopus_californicus_protein_v1.0|TCALIF_07877-PA
gml|Tigriopus_californicus_protein_v1.0|TCALIF_01298-PA
gml|Tigriopus_californicus_protein_v1.0|TCALIF_01297-PA
gml|Eurytemora_affinis_protein_v0.5.3|EAFP019734-PA
gml|Eurytemora_affinis_protein_v0.5.3|EAFP017339-PA
gml|Eurytemora_affinis_protein_v0.5.3|EAFP001603-PA
gml|Eurytemora_affinis_protein_v0.5.3|EAFP001528-PA
gml|Parasteatoda_tepidariorum_protein_v0.5.3|PTEP006966-PA
gml|Centruroides_exilicauda_protein_v0.5.3|CSCU000454-PA
gi|939674845|ref|XP_014298694.1|
gml|Trichogramma_pretiosum_protein_v0.5.3|TPRE008456-PA
gi|805810930|ref|XP_012147267.1|
gml|Agrilus_planipennis_protein_v0.5.3|APLA000706-PA
gml|Agrilus_planipennis_protein_v0.5.3|APLA003379-PA

gml|Eurytemora_affinis_protein_v0.5.3|EAFP027431-PA
gml|Tigriopus_californicus_protein_v1.0|TCALIF_07877-PA
gml|Tigriopus_californicus_protein_v1.0|TCALIF_01298-PA
gml|Tigriopus_californicus_protein_v1.0|TCALIF_01297-PA
gml|Eurytemora_affinis_protein_v0.5.3|EAFP019734-PA
gml|Eurytemora_affinis_protein_v0.5.3|EAFP017339-PA
gml|Eurytemora_affinis_protein_v0.5.3|EAFP001603-PA
gml|Eurytemora_affinis_protein_v0.5.3|EAFP001528-PA
gml|Parasteatoda_tepidariorum_protein_v0.5.3|PTEP006966-PA
gml|Centruroides_exilicauda_protein_v0.5.3|CSCU000454-PA
gi|939674845|ref|XP_014298694.1|
gml|Trichogramma_pretiosum_protein_v0.5.3|TPRE008456-PA
gi|805810930|ref|XP_012147267.1|
gml|Agrilus_planipennis_protein_v0.5.3|APLA000706-PA
gml|Agrilus_planipennis_protein_v0.5.3|APLA003379-PA
```

```
gml|Agrilus_planipennis_protein_v0.5.3|APLA000706-PA
gml|Agrilus_planipennis_protein_v0.5.3|APLA003379-PA
```

```
gml|Eurytemora_affinis_protein_v0.5.3|EAFP027431-PA
gml|Tigriopus_californicus_protein_v1.0|TCALIF_07877-PA
gml|Tigriopus_californicus_protein_v1.0|TCALIF_01298-PA
gml|Tigriopus_californicus_protein_v1.0|TCALIF_01297-PA
gml|Eurytemora_affinis_protein_v0.5.3|EAFP019734-PA
gml|Eurytemora_affinis_protein_v0.5.3|EAFP017339-PA
gml|Eurytemora_affinis_protein_v0.5.3|EAFP001603-PA
gml|Eurytemora_affinis_protein_v0.5.3|EAFP001528-PA
gml|Parasteatoda_tepidariorum_protein_v0.5.3|PTEP006966-PA
gml|Centruroides_exilicauda_protein_v0.5.3|CSCU000454-PA
gi|939674845|ref|XP_014298694.1|
gml|Trichogramma_pretiosum_protein_v0.5.3|TPRE008456-PA
gi|805810930|ref|XP_012147267.1|
gml|Agrilus_planipennis_protein_v0.5.3|APLA000706-PA
gml|Agrilus_planipennis_protein_v0.5.3|APLA003379-PA
```

uncolorful To hmmsearch



# HMMer

## HMMER

Tutorial

### Organisms

- ☐ All organisms
- ☒ *Agrilus planipennis*
- ☐ *Amyelois transitella*
- ☐ *Anoplophora glabripennis*
- ☐ *Athalia rosae*
- ☐ *Bactrocera cucurbitae*
- ☒ *Bactrocera dorsalis*
- ☐ *Bactrocera oleae*
- ☒ *Blattella germanica*
- ☐ *Cataglyphis aquilonaris*
- ☐ *Centruroides exilicauda*
- ☒ *Ceratitis capitata*
- ☐ *Cimex lectularius*

### Ceratitis capitata

#### Protein

- ☒ Protein - NCBI Predicted protein coding genes, Annotation Release 100
- ☒ Protein - Ceratitis capitata JAMG\_v1, peptides

### Query Sequence / Multiple sequence alignment

Your sequence is not detected as fasta (phmmmer disabled):

CLUSTAL O(1.2.0) multiple sequence alignment

```
gnl|Eurytemora_affinis_protein_v0.5.3|EAF027431-PA
MCDEDVAALVVDNGSGMCKAGFAGDDAPRAVFPSTVGRPRHQGVMVGMGQKDAYVGDEAQ
gnl|Tigriopus_californicus_protein_v1.0|TCALIF_07877-PA
```

## HMMER Success

### Download

Input file

Hmmer result

Submission Details

### Report Details

Jump To Dataset **baccuc\_v100\_protein.fa**

```
# hmmsearch :: search profile(s) against a sequence database
# HMMER 3.1b2 (February 2015); http://hmmer.org/
# Copyright (C) 2015 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
# -----
# query HMM file:          3d018b4f3bd74fbb9be95a2853b20987.in.hmm
# target sequence database: baccuc_v100_protein.fa
# output directed to file: 0.out
# sequence reporting threshold: E-value <= 0.01
# domain reporting threshold:  E-value <= 0.03
# domain inclusion threshold:  E-value <= 0.03
# -----
```

```
Query:          3d018b4f3bd74fbb9be95a2853b20987 [M=921]
Scores for complete sequences (score includes all domains):
--- full sequence ---  --- best 1 domain ---  -#dom-
E-value  score  bias  E-value  score  bias  exp  N  Sequence                                Description
-----
0 1674.8  22.6      0 1674.5  22.6    1.0  1  gi|751446905|ref|XP_011177564.1|  PREDICTED: N-alpha-acetyltransfe
0.0032  16.2   1.2    0.0032  16.2   1.2    2.9  3  gi|751456064|ref|XP_011182581.1|  PREDICTED: transmembrane and TPR
```

Domain annotation for each sequence (and alignment):



United States Department of Agriculture

'Frozen' genome  
assembly

Automated  
annotations

Ancillary datafiles (e.g.  
RNA-Seq alignments)

Submission

 **Workspace@NAL**  
<https://i5k.nal.usda.gov/>

## Resources

## Tools

## Services

**Organism  
Information  
Page**

**Custom BLAST  
interface**

**Manual annotation  
quality control**

**Official gene set generation**

**Bulk data  
downloads**

**JBrowse genome  
browser**

## Challenges

**Non-standard data  
formatting**

**Tutorials**

**Apollo manual  
curation tool**

**Failure to submit all metadata  
(ex: sample origin; analysis  
methods)**

**HMMer**

**Clustal**

# Post-Annotation QC

- Manual annotations are run through our Quality Control pipeline
- Some issues need manual intervention
  - Missing required fields
  - Complex splits/merges
  - Incomplete models and those abandoned in process
- Some issues can be automatically corrected
- Iterative process
  - Models requiring inspection are referred back to curators
  - After resolution models are screened again to screen for additional issues

# OGS Generation

- An Official Gene Set is the gene set chosen by the community to be the representative set of gene models for that organism
- Our system takes a single existing gene set and incorporates the validated manual annotations
- The gene set may be a previous OGS or other gene set (e.g. Maker models)
- Manual curations are used to
  - Update models
  - Flag models for removal from the final set
- The resulting set is then tested for errors and once approved, disseminated to the community

# Acknowledgements

- **The NAL team**

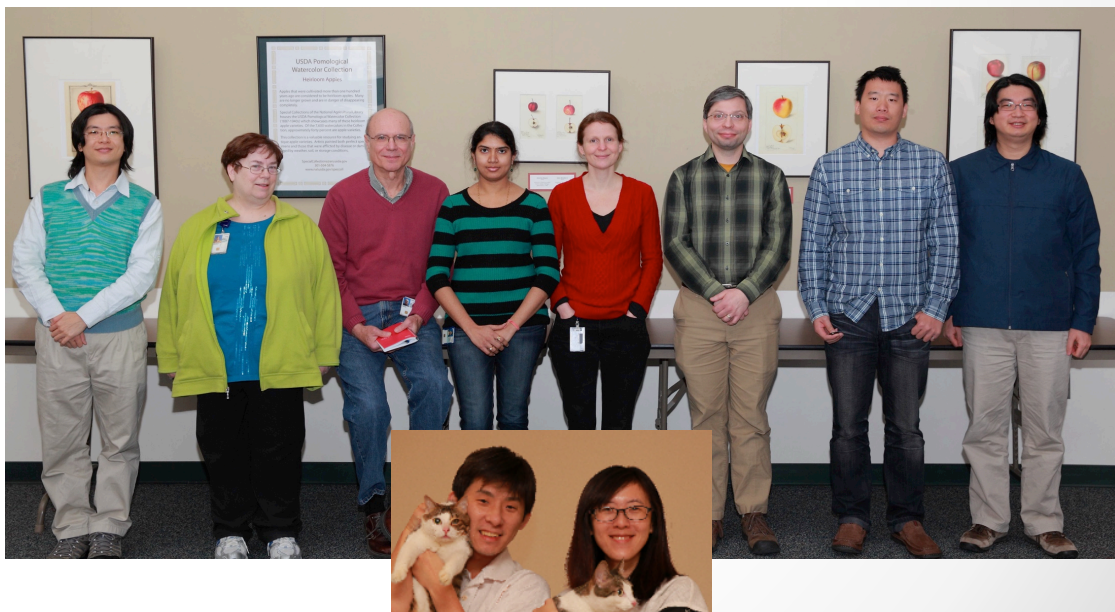
- Monica Poelchau
- Gary Moore
- Susan McCarthy
- Mei-Ju Chen
- Yu-Yu (Fish) Lin
- Limei Chiang
- Chao-I Tuan
- Chiatanya Gutta

- **i5k Workspace@NAL advisory committee**

- Jay Evans
- Don Gourley
- Kevin Hackett
- Simon Liu
- Ursula Pieper
- Paul Wester

- **Team Alumni**

- Chien-Yueh Lee
- Han Lin
- Jun-Wei Lin
- Vijaya Tsavatapalli





United States Department of Agriculture

# For More Information

Visit us!

<http://i5k.nal.usda.gov>

Contact us!

[i5k@ars.usda.gov](mailto:i5k@ars.usda.gov)

Check out our code on GitHub!

<https://github.com/NAL-i5K/>





**United States Department of Agriculture**

---



**United States Department of Agriculture**

---



**United States Department of Agriculture**

---